

Q1

There are two `atomicAdd` operations and one `atomicMin` operation for each input cell. There are three atomic operation for each input cell, which is total of $3 * \text{inputLength}$ operations.

Q2

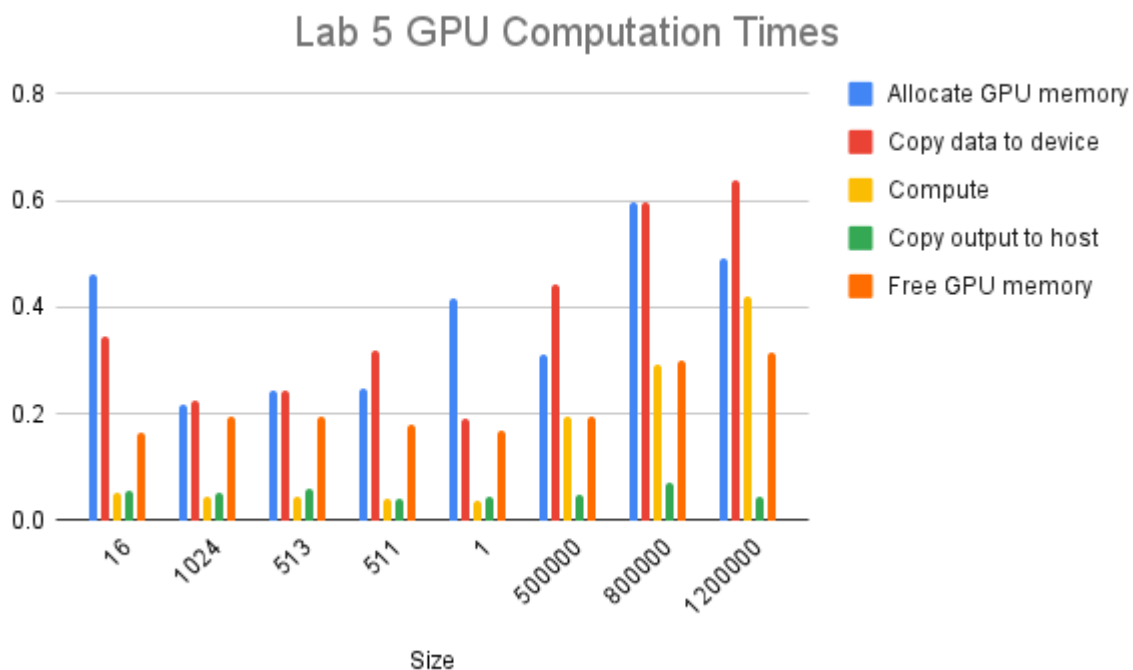
There will be a lot of contention to a shared memory address as every thread in a thread block will update the same location on the shared memory.

Q3

It would be unusual to have a contention as there are many possible values (4096) in the input which are random compared to the thread block size (1024).

Q5

The kernel was run on RTX 3060ti with 8GB of VRAM, Windows 11 WSL Ubuntu 22.04 with thread block size 1024.



The raw data table is as follows:

Size	Import data to host	Allocate GPU memory	Copy data to device	Compute	Copy output to host	Free GPU memory
16	0.065649	0.462526	0.346577	0.05075	0.057039	0.165257
1024	1.14765	0.216273	0.22418	0.045239	0.05364	0.194718
513	0.617615	0.244198	0.245348	0.044209	0.05841	0.196608
511	0.590085	0.246368	0.319857	0.04114	0.039659	0.179678
1	0.046059	0.416626	0.190209	0.03798	0.04633	0.169579
500000	545.169	0.310617	0.441416	0.195318	0.04729	0.194988
800000	876.457	0.596044	0.595325	0.291507	0.069709	0.300258
1200000	1279.9	0.490695	0.636415	0.418917	0.043499	0.316008